DOCUMENT RESUME

ED 430 046 TM 029 790

AUTHOR Sun, Anji; Schulz, E. Matthew
TITLE Evaluating College Services Using Student Ratings with Incomplete Data: An IRT Rating Scale Model Approach.
PUB DATE 1999-04-00
NOTE 32p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Students; Colleges; Comparative Analysis; Evaluation Methods; Higher Education; *Item Response Theory; *Rating Scales; Responses; *Student Personnel Services; Surveys
IDENTIFIERS *Missing Data

ABSTRACT

Missing data create problems for the interpretations and inferences of survey results, especially if the amount of missing data is substantial. This study considers issues of missing data when comparisons are made across survey items. It is suggested that if the groups responding to different items differ in their tendency to use one end of the measurement scale or the other, named as "pleasability" according to the specific nature of the data used in the study, then the comparisons of these items should be adjusted for this difference. Data for the study were obtained from the Student Opinion Survey processing history files of ACT Inc. for 10 institutions. Students at these colleges had responded to the survey about college services used, but the average number of students responding to all items was less than half the total number of respondents. The results of the study show a consistent difference in pleasability among the respondents and that an item response theory rating scale model can improve comparisons among survey items when the data consist of ratings on a Likert scale and respondents rate only selected items. (Contains 4 figures, 11 tables, and 17 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 430 046

Evaluating College Services Using Student Ratings with Incomplete Data: An IRT Rating Scale Model Approach

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Anji Sun
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

by

Anji Sun
and
E. Mathew Schulz

ACT, Inc.

prepared for
The Annual Meeting of American Educational Research Association
Montreal, Canada
April 1999

TM029790

BEST COPY AVAILABLE

Abstract

Missing data create problems for the interpretations and inferences of survey results, especially if the amount of missing data is substantial. In this study, the authors are concerned with issues of missing data when comparisons are made across survey items. The authors believe that if the groups responding to different items differ in their tendency to use one end of the measurement scale or the other, named as "pleasability" according to the specific nature of the data used in the study, then the comparisons of these items should be adjusted for this difference. The results of the study show a consistent difference in pleasability among the respondents and that an IRT rating scale model can improve comparisons among survey items when the data consist of ratings on a Likert scale and respondents rate only selected items.

In this study, we are concerned with the problem of missing data in surveys that ask respondents to rate programs, services, products, or individuals on a Likert scale. Respondents to such surveys are typically instructed to rate only the items (services, products, or individuals) with which they have had experience. These instructions, while reasonable, inevitably yield a data matrix with many missing responses. It is not uncommon to find institutional reports, newspaper accounts, or research papers in which the Likert-scale means of survey items are reported, and the items are compared or rank ordered by their mean ratings.

The problem is that missing Likert ratings may not be missing at random. This point was illustrated by Brady (1989) who imagined two types of people--cynics and Pollyannas. Cynics don't like to provide positive assessments for anything and therefore tend to confine their ratings to the lower range of the Likert scale, e.g., "very dissatisfied." Pollyannas see good in everything, and therefore confine their ratings to the upper range of the scale, e.g., "very satisfied." These tendencies exist regardless of the true performance of whatever they are rating. Now, if cynics are more likely than Pollyannas to have experience with a given survey item, that item will be disadvantaged when its mean rating is compared to the mean rating given to other items.

The missing data problems of survey research have been extensively studied, and a number of procedures that compensate for non-responses have been developed over the years (Kalton, 1983; Little and Rubin, 1987). Two general approaches to treat missing data are available-case method (listwise or pairwise deletion) and data imputations. A common conclusion is that if data are randomly missing, and if the amount of missing data is not excessive, then any treatment is as good as any other (Kromrey & Hines, 1994). However, if data are not missing at random, then the use of either method is questionable (Cohen & Cohen, 1983; Graham, Hofer, & MacKinnon, 1996; Kolb & Dayton, 1996; Kromrey & Hines, 1994; Witta, 1994).

To improve the precision of estimation when a large amount of non-random missing data is present, a group of maximum likelihood approaches has been developed that provides more accurate estimates than the deletion and imputation methods (Gross, 1990; Kolb & Dayton, 1996; Little & Rubin, 1989; Muthén & Kaplan, 1987). However, the mathematical complexity of these approaches and the unavailability of computer software for their implementation have prevented its common application in the field (Gross, 1990; Kromrey & Hines, 1994; Little, 1992).

In this study, we use item response theory (IRT) to compare items from a Likert survey when large amounts of data are missing. According to IRT, a Likert rating is a stochastic variable whose distribution is controlled by a person parameter and one or more item parameters. If the rating scale ranges from "very dissatisfied" to "very satisfied," the trait measured by the person parameter might be called "pleasability." Persons with low pleasability tend to use the lower end of the rating scale (very dissatisfied), while persons with high pleasability tend to use the higher end of the rating scale (very satisfied). By measuring these tendencies, IRT controls for their effects on the item parameters. This means that the proportion of cynics to Pollyannas can vary across items without affecting the parameters of the items. In this situation, the item

parameters in the IRT model are comparable, but the simple means of ratings given to the items are not.

The hypotheses in this study broadly state conditions that we believe are generally met in rating scale data from surveys. In terms of the trait, pleasability, these hypotheses are:

1) the survey items measure reliable differences in the pleasability of survey respondents,
2) the groups rating different items are not equivalent in pleasability,
3) comparisons of items through IRT are better than comparisons through simple means, and
4) the data for each survey item have acceptable fit to the IRT model.

Hypothesis 1 follows from the general notion that people differ in whether they tend to use the high or low end of a Likert rating scale and that this difference is expressed consistently across the items to which the Likert scale is applied. In effect, all of the items work together in a unidimensional fashion to measure this tendency. When Likert scales are used to measure attitudes, and in other contexts where the focus is on comparisons of persons with respect to a trait of interest, this hypothesis may be considered too obvious to present formally. It amounts to saying that one can define a reliable measure of the trait. In survey work where the focus is on comparing services or products, the idea that a reliable person-trait might be measurable from the same data might seem surprising.

Hypothesis 2 is worth considering if Hypothesis 1 is true and there is a large amount of missing data in the survey. Consider, for example, a case where 50 persons respond to item A, 100 persons respond to item B, and only 15 persons are in both groups. The 35 persons who responded to item A only could differ significantly in pleasability from the 85 persons who responded to item B only. If pleasability affects ratings irrespective of the true performance of items, items are comparable through their simple mean ratings only if the A-only group is randomly equivalent in pleasability to the B-only group. If the groups are not equivalent in pleasability, a fair comparison of one item to another can occur only through the persons who rated *both* items (i.e., N=15 persons for items A and B) or by using a special analysis, such as IRT that controls for pleasability differences between groups.

Assuming Hypotheses 1 and 2 are correct, Hypotheses 3 states that IRT-based comparisons among survey items are demonstrably better than comparisons based on simple means. In order to demonstrate this, IRT model parameter-estimates will be used to generate an IRT-predicted mean rating for each item. The IRT-predicted mean rating is the "simple"mean we would expect for the item if *all* the survey respondents had rated the item. Simple means, and IRT-predicted means will be evaluated for each pair of survey items. For example, mean ratings for items A and B (see preceding paragraph) will be computed using the fifteen persons who rated *both* items--these *matched-group* means will be the criterion. The simple means and the IRT-predicted means for items A and B should agree with the matched-group means with regard to 1) which item is more satisfactory and 2) approximately how much more satisfactory one item is than another. It is, thus, agreement with the *difference* between matched-group means that we expect the IRT-predicted means to improve upon in comparison to simple means.

The gist of Hypothesis 4 is that the tendency of persons to use lower ends of the rating scale (e.g. cynics) or higher ends (Pollyannas) will not vary substantially among items. The IRT model predicts that cynics, for example, tend to rate *all* items lower than Pollyanna's. If there is one item that does not conform to this prediction, then IRT-based comparisons should not be made for that item. (IRT-based comparisons for other items may still be made.)

Method

Data description

Data for the present study were obtained from *Student Opinion Survey* (SOS) processing history files maintained by ACT, Inc. Section II of the SOS lists 23 services provided by colleges. (See the "item content column" of Table 3.) The services include day care, parking facilities, veteran's services, advising, and student counseling. Students are asked to indicate whether they have used the service. If yes, they are asked to rate their satisfaction with the service on a five-category rating scale ranging from "very satisfied" (1) to "very dissatisfied (5). Fifty-seven post-secondary education institutions administered the SOS survey in 1998.

Ten of these institutions were selected for this study based on the number of students who responded to the SOS (300 or more). No attempt was made to control for any of the characteristics of the institutions (e.g., public/private affiliation, location, enrollment, etc.). The total number of respondents per school and average sample sizes per item and person are shown in Table 1. Schools are identified by the numbers 1 to 10. The number of students per institution ranged from 376 to 1358.

The extent of missing data in this study is indicated by the "Item sample size" and "Person sample size" information in Table 1. Item sample size is to the number of items responded to by a given person, and persons sample size is the number of persons who rated a given item. In over half the schools, no person responded to all of the items (maximum item-sample size is less than 23), and in almost all schools there was at least one person who rated only one item (minimum sample size is 1). The average number of items rated per respondent (item sample size) was no greater than 12 (School 5), and was as small as 6.8 (Schools 8 and 9).

In all schools, the average number of persons responding to an item was less than half the total number of respondents. In six schools, at least one item was rated by five or fewer persons, and in one school an item was rated by only one person. In every school, there was at least one item (usually item 6: library facilities) that was rated by nearly all of the respondents, but no item was ever rated by all respondents.

Insert Table 1 about here

The percentage of ratings for each item across the ten schools is shown in Table 2. Item 6 (library facilities) was rated by the largest percentage of respondents (83.6% on average). Item 23 (Day care services) was rated by the lowest percentage of respondents (1.2%) on average. Six

of the 23 items were rated by fewer than five percent of the respondents in one or more schools. (See "minimum"column). These were student health insurance (item 8), resident hall services and programs (item 12), credit-by-examination programs (item 17), college mass transit services (item 20), veterans services (item 22), and day care services (item 23).

Insert Table 2 about here

Simple Mean Analysis

Mean ratings per service were computed within schools. Sample size per service was the number of persons within each school who had experience with and rated the service. Items were ranked by their mean ratings. The results of the simple mean analysis for one of the schools (School 9) are shown in Table 3. Person sample size ranges from 2 (item 23: day care services) to 355 (item 6: library facilities). Simple means on the 5-category rating scale (very dissatisfied = 1, very satisfied = 5) ranged from 4.6 for item 22, veterans services (rank = 1), to 2.05 for item 21, parking facilities and services (rank = 23).

Insert Table 3 about here

IRT-analysis

The rating scale data were analyzed separately for each institution with the computer program Bigsteps (Wright and Linacre, 1991). The rating scale model (Andrich, 1978a,b) was used for the analysis. The rating scale model for item *i* includes a combination of one unique item parameter, D*i*, and a series of step parameters, called thresholds, that are the same for all items. The step 1 threshold, for example, represents the relative amount of pleasability that it takes to prefer the response, "dissatisfied," to the response, "very dissatisfied," for any item. The step-2 threshold represents the relative amount pleasability that it takes to prefer the "neutral" response to the "dissatisfied" response for any item. The number of step thresholds, *m*, is one less than the number of categories in the rating scale.

One formulation of the rating scale model is:

$$\log\left[P_{nix} / P_{ni(x-1)}\right] = B_n - D_i - F_x \qquad x = 1,\ldots,m \qquad (1)$$

where the rating scale categories are assigned the integers 0,1,…,m, and

P_{nix} is the probability that person *n* responds in category *x* to item i,
$P_{ni(x-1)}$ is the probability that person *n* gives a rating of *x*-1 to item *i*,
B_n is the pleasability of person *n*,
D_i is the difficulty of item *i*, and
F_x is the threshold parameter of step x.

Higher values of B_n mean that person *n* is more pleasable. Higher values of D_i mean that item *i* is *less* pleasing or is more "difficult" for persons to feel satisfied towards. On the original rating scale metric with "very dissatisfied" = 1 and "very satisfied" = 5, the expected rating of item *i* by person *n* is:

$$E_{ni} = 1 + \sum_{x=1}^{m} x * P_{nix} \qquad (2)$$

Joint maximum likelihood estimates of the model parameters can be obtained in the presence of missing data by summing for each person, E_{ni} over only the items that they responded to, and summing for each item, E_{ni} over only the persons who rated the item. This is done iteratively with adjustments being made to item and person parameter estimates between iterations in order to close the gap between expected and observed total scores associated with each item and person.

An "extreme"person in this analysis is one who chooses "very dissatisfied" for every item or "very satisfied"for every item. Maximum likelihood measures cannot be obtained for these persons, but were supplied through an optional feature of the Bigsteps program that involves certain reasonable assumptions.

A complete data matrix of expected scores was generated from the parameter estimates for all persons, items, and step thresholds. All measured persons, including measures supplied by default for extreme persons were included. The expected score resulting from a person and an item was obtained via equation 2, where

$$P_{nix} = \frac{\exp \sum_{j=0}^{x} [B_n - D_i - F_j]}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} [B_n - D_i - F_j]} \qquad x = 1,\ldots,m \qquad (3)$$

and $P_{ni0} = 1 - [\Sigma(P_{nix}), x=1,\ldots,m]$.

The IRT-predicted mean rating for an item was the average expected score over all persons (including extreme persons). Items were then ranked according to their IRT-predicted mean ratings.

Evaluation of the Hypotheses

Hypothesis 1 was evaluated through the estimation of reliability of person separation on the measure of pleasability, which is routinely produced by the Bigsteps program. These estimates are computed as one minus the ratio of the mean squared error of person measures to the variance of the person measures. They are comparable in magnitude to Cronbach's alpha coefficient.

Hypotheses 2 and 3 were evaluated through secondary data analyses based on pairs of items. With 23 items, there were 253 possible item-pairs within each school. The critical

information extracted for each pair of items is exemplified in Table 4. This information pertains to school 9 data for items 4 and 5. Table 4 shows that 92 persons rated item 4, 84 persons rated item 5, and only 33 persons rated both items. According to persons who rated both items, item 4 is .21 rating scale units more satisfactory than item 5. According to each item's total group (simple means), item 5 is .11 rating scale units more satisfactory than item 4. This disagreement can be seen to stem from the groups who rated one item, but not the other. According to these disjoint groups, item 5 is .29 rating scale units more satisfactory than item 4. An explanation for this discrepancy is seen in the mean pleasability measures of the disjoint groups. The group that rated item 5 only is more pleasable (mean = 1.22) than the group that rated item 4 only (mean = .57).

Evaluation of Hypothesis 2. Analyses of variance (ANOVA) were performed on the pleasability measures of the disjoint groups within each item pair. A separate analysis was performed for each item pair within each school. The number of ANOVAs performed per school approached or equaled (in some cases) the number of possible item pairs (253). The results for each school are summarized as the percentage of ANOVAs for which the difference between disjoint groups is statistically significant at the .05 level. If this percentage is above 10 for a given school, we will consider this as a strong indication that the groups responding to each item are not randomly equivalent within that school. If this percentage is reached for half of the schools, we will consider this as a strong indication that groups responding to different items on our survey are, in general, not randomly equivalent.

Insert Table 4 about here

Evaluation of Hypothesis 3. Various estimates of the *difference* between two items in rating scale units are shown in the last row of Table 4. Let D_{simple} denote the difference based on simple means (total number of persons rating each item), $D_{matched}$ denote the difference based on matched group means (persons who rated both items), and D_{irt} denote the difference based on IRT predicted means (all measured respondents). The following variables are defined in terms of D_{simple}, $D_{matched}$, and D_{irt}:

δ_{simple} the difference between D_{simple} and $D_{matched}$ (D_{simple} - $D_{matched}$)

δ_{irt} the difference between D_{irt} and $D_{matched}$ (D_{irt} - $D_{matched}$)

σ_{simple} the standard deviation of δ_{simple} across item pairs and schools,

σ_{irt} the standard deviation of δ_{irt} across item pairs and schools,

ρ_{simple} the correlation between D_{simple} and $D_{matched}$ across items and schools,

ρ_{irt} the corresponding correlation between D_{irt} and $D_{matched}$.

π_{simple} the proportion of item pairs exhibiting sign disagreement between D_{simple} and $D_{matched}$

π_{irt} the proportion of sign disagreement between D_{irt} and $D_{matched}$.

Estimates of these quantities were obtained using the PROC MEANS and PROC CORR procedures in SAS. Different sets of estimates were obtained depending upon the number of persons in the "matched"group for an item pair (persons rating both items within a pair). One set of estimates was computed using all available item pairs (N=2493 combined across schools), which included those with as few as one person rating both items. Estimates were then based on increasingly larger sample size ranges in order to assess the effect of person-sample size on the need for using IRT-predicted means rather than simple means in this kind of work.

Hypothesis 3 was evaluated in terms of three expectations.

Expectation 1: $\sigma_{irt} < \sigma_{simple}$.

Expectation 2: $\rho_{irt} > \rho_{simple}$

Expectation 3: $\pi_{irt} < \pi_{simple}$

No formal statistical tests were used to confirm the first two expectations. If Hypotheses 1, 2, and 3 were true and data fit the IRT model, we could not expect D_{simple}, $D_{matched}$, and D_{irt} to be equal. This is because the groups that are used to define these quantities would vary in pleasability, as can be seen in Table 4. This means that the discrepancies, δ_{simple} and δ_{irt}, may not have a zero mean and strictly equal variance under Hypothesis 3. Also, ρ_{irt} and ρ_{simple} may have different expected values less than 1.0. We do not believe that comparison of these quantities will be unproductive just because they do not conform to the strict assumptions of statistical hypothesis testing.

A formal test of Expectation 3 is supported because $D_{matched}$, and D_{irt} should have the same sign, even if they are not necessarily equal, under the hypotheses in this study. This is because the relationship between person's pleasability and the expected score on an item is monotonic according to the IRT model. Also, according to the model, D_{simple} and $D_{matched}$ should occasionally have opposite sign, if difference in the pleasability of the groups rating the items is large enough. Expectation 3 was tested using the normal approximation to the binomial distribution.

Evaluation of Hypothesis 4. Item fit was assessed by the mean squared residual (outfit) statistic. This statistic is routinely computed by the Bigsteps program. The misfit statistic has an expected value of 1.0 when data fit the model. Values less than 0.6 or greater than 1.4 were flagged for closer inspection.

To assess the practical consequences of item misfit, a secondary analysis was performed on the data. All measured persons (including extreme persons) were divided into three equal-sized groups within each school according to their measure of pleasability. For easy reference,

persons within these groups are called "cynics," "neutrals," or "Pollyannas." The location of group boundaries on the pleasability scale varied across schools so that the group sizes would be equal within schools.

For each item, IRT-predicted and observed means were obtained using persons within each pleasability group who responded to the item. For example, a school with 600 total measured respondents would have 200 persons in each group, but the sample sizes for a particular item might be 60, 50, and 40 (respectively for cynics, neutrals, and Pollyannas.) The IRT-predicted and observed mean rating of the 60 cynics was compared to each other and to those of the 50 neutrals and 40 Pollyannas.

In order for the IRT-predicted overall mean rating to be preferred over the simple mean rating for a given item, it is important that the observed mean ratings within pleasability group for the item conform to IRT-predictions. If cynics tend to rate an item as highly as Pollyannas, then a correction for pleasability is not needed for that item. Items of this type are expected to "underfit" the IRT model, i.e., to have misfit statistics substantially greater than 1.0. On the other hand, a tendency of cynics to rate an item even worse, in comparison to Pollyannas, than that predicted by the IRT model means the item "overfits" the model. Items of this type should have indices of misfit substantially less than 1.0 and need more correction than provided by the IRT-predicted mean rating.

Results

Hypothesis 1: Respondents vary in pleasability

The distribution of pleasability measures from School 9 is shown in Figure 1. Other schools in this study had similar distributions. Pleasability measures in School 9 range from a low of -2.6 (very difficult to please) to a high of 4.6 (very easy to please). The estimated reliability of pleasability measures in School 9 is .66.

Insert Figure 1 about here

The relationship between the reliability of the person measures within a school and the average item-sample size per person is shown in Figure 2. The average item sample size is the average "test length." Variation in test length accounts for much of the between-school differences in the reliability of pleasability measures. The lowest reliability, .58, comes from the school having the second shortest average test length (7 items). The highest reliability, .79, comes from the school having the longest average test length (12 items). The efficiency of these survey items in measuring pleasability is comparable to the efficiency of instruments that are expressly designed for measuring person traits such attitudes.

Insert Figure 2 about here

Hypothesis 2: Groups responding to different items differ in pleasability

Results of ANOVAs performed on the pleasability measures of disjoint (non-overlapping) groups responding to each item within a pair are shown in Table 5. The number of item pairs on which such an ANOVA could be performed ranged from 200 (School 7) to the maximum possible number, 253 (Schools 4 and 8). For some item pairs, sample sizes in one or more of the disjoint groups were very small (as small as 1), resulting in low statistical power to detect differences.

Nevertheless, in eight of the schools, over ten percent of the ANOVAs yielded p-values of less than .05 for the null hypothesis of no difference between disjoint groups. The highest percentage was 31 (School 8). And the two lowest percentages were 8 (School 7) and 9 (School 6). Although the ANOVAs are not strictly independent due to the certain presence of the same person(s) in more than one ANOVA, these results are a strong indication that the groups responding to different items are not equivalent in pleasability.

Insert Table 5 about here

Hypothesis 3: IRT-based comparisons are better than comparisons through simple means

Results of comparing simple and IRT-predicted mean to the matched-means within item-pairs are shown in Table 6. Recall that the matched mean *difference* is the criterion and is derived from persons who rated *both* items in the pair. Results are combined across schools but are presented according to the number of persons who rated both items within an item pair. Different rows in the table correspond to different sample sizes.

The results for all pairs combined are shown in the first row of Table 6, with row heading "≥1". Out of 2530 possible item pairs (10 schools times 23-choose-2), there were 2493 in which at least one person rated both items in the pair. All expectations expressed with respect to Hypothesis 3 are confirmed by these data. Specifically, compared to the simple mean difference, the IRT-predicted mean difference 1) is closer to the criterion (σ_{irt} (.31) < σ_{simple} (.40)); 2) exhibits less sign-disagreement with the criterion (π_{irt} (8%) < π_{simple} (10%)); and 3) has a higher correlation with the criterion (ρ_{irt} (.88) > ρ_{simple} (.82)). The difference between π_{irt} (8%) and π_{simple} (10%) is statistically significant at the .01 level.

The remaining rows in Table 6 show that the improvement of IRT-predicted means over simple means is greatest when there are few persons who rate both items (matched group), and diminishes with increasing numbers in the matched group. With only 2 to 9 people in the matched group, the IRT-predicted mean difference has substantially higher correlation with the criterion (.64 versus .54), has much less sign disagreement (13% versus 18%), and is much closer to the criterion (.55 versus .65 for the expected root mean discrepancy) compared to the simple mean difference. The difference in sign disagreement (13% versus 18%) does not reach statistical significance due to the smaller number of item pairs involved in this result (373).

Differences in sign disagreement listed in the remaining rows of Table 6 are also not statistically significant for the same reason.

With 500 or more persons rating both items, the improvement of IRT-predicted means over simple means is relatively slight. Due to rounding, the only improvement detectable in Table 6 is a one percentage-point decrease in sign disagreement (1% versus 2%).

Insert Table 6 about here

IRT item-rankings are compared with simple-mean item-rankings for School 9 in Table 7. Shown in the last column (Column 10: Change in Rank) is the difference between the rank order of the item's performance according to the IRT (Column 8) and the item's rank order according to simple means (Column 9). At the extremes (first and last rows), the IRT analysis improves the rank of Item 4, "job placement services" by 6 positions (from position 16 to position 10) and decreases the rank of Item 23, "day care services," by 17 positions (from position 2 to position 19).

The mechanism for these and other changes is also presented in Table 7. The items are sorted according to the average pleasability of the persons rating the items. The persons rating Item 4 were the least pleasable group (N = 92, mean pleasability = .62) while the persons rating Item 23 were the most pleasable group (N=2, mean pleasability = 4.0). The group differences in pleasability shown in column 4 of Table 7 largely account for the difference between the IRT-predicted mean rating (column 5) and the simple mean rating (column 6). This difference is shown in column 7 as the "Change in Mean." The correlation between mean pleasability of persons rating the item (column 4) and the change in the item mean (column 7) is .-.96 in School 9. The coefficient for this correlation is similarly strong in other schools, e.g., -.98 in School 8 and -.88 in School 10.

insert Table 7 about here

The differences between IRT results and simple mean results across schools are summarized in Tables 8 and 9. With 10 schools and 23 items, there are 230 possible changes to account for. Change in rank is summarized in Table 8 as the frequency of various magnitudes of absolute change. Of 230 possible changes, there was no change in 78 (34%) and a change of only 1 or 2 positions in 80 (35% of the possibilities). Thus, only about 30% of the cases exhibited a change in rank of more than two positions. Six percent of the cases exhibited a change in rank of 6 or more positions (combining the last two rows of Table 8).

Insert Table 8 about here

A frequency distribution based on amount of absolute change in the mean rating is presented in Table 9. Out of 230 possible changes, there was no change (rounded to nearest 0.1) in 132 (57.4%) and a change of only 0.1 in another 61 (26.5%). Combining the remaining rows of this table, change of 0.2 (absolute value) or more in the mean rating was exhibited in 37 instances, which is about 16% of the possible number.

Insert Table 9 about here

Hypothesis 4: Items show good individual fit to the IRT model

A histogram of the mean squared residual (outfit) statistics combined across items and schools (N=230) is shown in Figure 3. The mode of the distribution is 1.0, as expected. Relatively few item outfit statistics fell outside the range of 0.6 to 1.4. Mean squared residuals within this range are generally considered acceptable. Eight outfit statistics were below 0.6 and eight were above 1.4.

Insert Figure 3 about here

Detailed information about the cases of underfit (outfit > 1.4) and overfit (outfit < 0.6) are displayed in Table 10. Eleven of the sixteen cases involved the last three items on the survey. Item 21 was involved in two cases of underfit (Schools 1 and 9). Item 22 was involved in two cases of overfit (Schools 4 and 5) and two cases of underfit (Schools 2 and 8). Item 23 was involved in three cases of overfit (Schools 3, 5, and 9). These items also tend to be associated with small sample sizes. Sample size is less than fifteen in six of the eleven cases involving these items.

The mean rating by pleasability group is shown for each item in Table 10. Mean ratings by group are expected to increase from cynics to neutrals to pollyannas. With sixteen cases of misfit and three pleasability groups, data in this table allow thirty-two comparisons of means against this expectation. A decrease is evident in only three comparisons. Two of these decreases involved Item 23 (Schools 3 and 5) and were associated with *over*fit! (Overfitting items should have a greater tendency to show *increases*.) The other decrease involved item 21 in School 9, where the mean rating by cynics (1.7, N=20) exceeds the mean rating by neutrals (1.6, N=27).

The two exceptions involving Item 23 appear to be due to the extreme ratings given to Item 23 (overall means of 3.4 and 3.6 respectively in Schools 5 and 9). Extreme items tend to be associated with overfit due to lack of variation in the ratings people choose for these items. One would not expect much variation in the means of pleasability groups to such items. The exceptional differences between pleasability group means for Item 23 are in fact, quite small and the pleasability group sample sizes for these means are also small.

Insert Table 10 about here

A plot of IRT-predicted and observed means by pleasability group for a case of overfit (0.4 for Item 12, School 9) and underfit (1.5 for Item 21, School 1) is shown in Figure 4. This plot shows that, for both underfitting and overfitting items, observed mean ratings strongly tend to increase with pleasability. With overfitting items, the trend is stronger than predicted by the IRT model. With underfitting items, the trend is not as strong as predicted, but is still strong.

As a final check on how pervasive this trend is in our data, we counted the total number of cases, out of 230 possible, in which mean rating failed to increase with pleasability group. This failure was observed in only six cases including the three displayed in Table 10. These cases, like those in Table 10, involved small sample sizes for one or both of the means exhibiting the exception to the trend.

Insert Figure 4 about here

Finally, information related to the fit and utility of the IRT analysis for Item 4, "job placement services" is shown in Table 11. The changes in rank for Item 4 were among the largest and most consistent across schools. This can be seen in the last column in this table. In nine cases, the IRT analysis increased the performance rank of Item 4. (A negative change means the rank increased.)

The reasons for this consistent change in rank are clear from a comparison of the sample sizes and mean ratings of pleasability groups for Item 4 within schools. First, cynics were more likely than Pollyannas to have experience with this service. In School 9, for example, Item 4 was rated by 44 cynics, 31 neutrals, and 17 Pollyannas. Second, in every school the differences between pleasability group means were large and in the expected direction. This is why, for example, the mean pleasability of persons rating Item 4 in School 9 was the lowest (.62) for any item. (See column 4, Table 7.)

As shown in the second-to-last column of Table 11, the fit statistics for Item 4 ranged from a low of .88 (School 2) to a high of 1.37 (School 1). All of these fit values are within the acceptable range of .6 to 1.4. Thus, one would conclude from these results that the change in Item 4's ranking in the IRT analysis is reasonable in every school.

Insert Table 11 about here

Discussion

In this paper, we have shown that an IRT analysis can improve comparisons among survey items when the data consist of ratings on a Likert scale and respondents rate only the items with which they have had experience. One of the conditions for the specific improvements we've demonstrated is that there must be a lot of missing data and that some items will not be rated by many of the same respondents. This problem was evident in our data and is a general problem in surveys of this nature.

We then showed that respondents differed in their tendency to use one end of the rating scale or the other. Due to the specific nature of the rating scale in this study, we characterized this tendency as "pleasability." The reliability of the pleasability measures was in line with the reliability of other traits, such as attitudes, when measured by comparable numbers of rating-scale items. This is surprising because the focus in our survey, as in many others, is not on measuring respondents in any sense with the items, but in making comparisons among items.

Next, we showed that the person-samples for different items varied substantially in pleasability. This is a problem mostly for items with small sample size. We showed that mean ratings of items differ partly because of differences in the pleasability of persons rating the items. We believe this is undesirable and that any comparison between any two items should be controlled for differences in pleasability and should agree with a comparison that is based strictly on persons who rate both items.

Using item-pairwise comparisons we showed that IRT-results, compared to simple means, were more in agreement with comparisons where the same persons rated both items. This seems to us a good demonstration of the need to adjust item comparisons for differences in the pleasability of persons rating the item, and of the utility of an IRT analysis to make this adjustment.

The analysis of the fit of data to the model focused on items because the focus of the survey is on items. Individual item fit statistics were studied because it is a real possibility that IRT-results might be used to make decisions about some items (items fitting the model), but not others (misfitting items). Because the IRT-adjustment is strongly related to the modeled trend of higher ratings with higher pleasability, the relationship of the fit statistic to this trend was of primary interest. Responses to virtually all items exhibited this trend quite strongly. The fit statistics were very sensitive to whether items exhibited this trend more or less strongly than expected by the model. We conclude that the IRT-adjusted means are better than the simple means for every item in this study.

Conclusions about items with very small sample sizes, such as 2 raters for item 23 in School 9, are highly tentative whether they are based on IRT parameter estimates or simple means. We recommend that minimum sample size requirements be set for either case, but see no reason why they should be higher for IRT analyses than for simple means if the data fit the IRT model reasonably well. Perhaps a better recommendation than setting minimum sample size requirements, is that the quality of the data for each item should be assessed routinely. For example, the item fit analysis revealed various problems that would affect any inference about

the item, not just those based on IRT. For example, it was evident that item 23 in school 9 was rated by persons who either responded only to that item or indicated they were "very satisfied" with every item they rated.

We believe that further research in this area could demonstrate the advantages of evaluating survey items in terms of their IRT parameter estimates *in addition to* rank order and mean rating comparisons. As shown in this study, IRT parameter estimates can be translated into expected mean ratings. Mean ratings (based on a common group or reference for all items) are useful for conveying to the layman where the item stands with respondents in terms of the rating scale. It is also occasionally useful to simplify comparisons among items by looking at ranks. However, we feel it is also useful to maintain a framework that quantifies "how much" better one item performs than another more on an equal-interval scale. If this were true, then it would make sense to use IRT not just as a tool for solving a missing data problem, but as a more general tool for making comparisons among survey items.

One criticism we anticipate is that people become cynics or Pollyannas as a *result* of their experience with these survey items and that it is incorrect to "adjust" for a disposition that is caused by the item. For example, item 4, "job placement services," may be so bad that any person coming into contact with this service tends to turn into a cynic. Why then should we "adjust" for the fact that the persons who rated this item tended to be cynics? The answer for this example is that the overall standing of item 4 with respect to other items is relatively good. It seems unlikely that it could be the cause of cynicism among the respondents.

It seems more likely that pleasability, as defined in this study, is a trait that is expressed uniformly with respect to college services. "Cynics" in our study are cynics only in relation to this set of survey items (and other items in a domain that these survey items could be said to represent). But this is true for attitude measures as well. A positive or negative attitude is defined only with respect to a specific object or domain of items. Cynics in this study are cynics towards college services, but perhaps not cynics generally.

Finally, the mechanism by which persons "self select" the services they rate in a survey needs to be distinguished from what goes on in achievement testing. In achievement testing, examinees are motivated to achieve a high score. Items are designed to measure maximal performance. If allowed to choose which items to answer, examinees will choose items non-randomly with respect to the probability of their score on the items. In survey work, respondents are not motivated to achieve a "high score." Items are not designed to elicit a maximal response, but rather a typical response. Respondents choose items based on whether or not they have experience with it, not on how high or low the respondent feels he/she is likely to rate the item. Survey respondents can therefore be said to answer an item randomly with respect to the probability of the rating they give to the item. This makes an IRT analysis appropriate for handling missing data in a survey, although it may not solve the problem of missing data in achievement testing.

Reference

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrica, 43*, 561-573.

Andrich, D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38*, 665-680.

Brady, H. E. (1989). Factor and ideal point analysis for interpersonally incomparable data. *Psychometrika, 54*(2), 181-202.

Braun, H. I. (1988). A new approach to avoiding problems of scale in interpreting trends in mental measurement data. *Journal of Educational Measurement, 25*(3), 171-191.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31*(2), 197-218.

Gross, A. L. (1990). A maximum likelihood approach to test validation with missing and censored dependent variables. *Psychometrika, 55*(3) 533-549.

Kalton, G. (1983). *Compensating for missing survey data*. (Research Report Series, Institute for Social Research) Ann Arbor, MI: The University of Michigan.

Kolb, R. R., & Dayton, C. M. (1996). Correcting for Nonresponse in Latent Class Analysis, *Multivariate Behavioral Research, 31*(1), 7-32.

Kromrey, J. D. & Hines, C. V. (1994). Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments. *Educational and Psychological Measurement, 54*(3), 573-593.

Little, R. J. (1992). Regression with missing *x*'s: A review. *Journal of the American Statistical Association*, 87, 1227-1237.

Little, R. J., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley and Sons.

Little, R. J., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research, 18*, 292-326.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*(3), 431-462.

Witta, E. L. (1994). Are values missing randomly in survey research? Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Nashville, TN.

Wright, B. D. & Linacre, J. M. (1991). BIGSTEPS. [A Rasch-model computer program.] MESA Press: Chicago, IL.

Zwick, R. (1992). Statistical and psychometric issues in the measurement of educational achievement trends: Examples from the National Assessment of Educational Progress. *Journal of Educational Statistics*, *17*(2), 205-218

Table 1.
Average Number of Items Rated and Average Number of Persons Rating Any One Item with School

School		Item Sample Size[a]			Person Sample Size[b]		
Id	n	Mean	Minimum	Maximum	Mean	Minimum	Maximum
1	376	8.8	1	18	124	2	310
2	672	9.0	1	20	243	13	604
3	718	9.4	1	23	261	5	601
4	1347	10.7	1	22	606	10	1254
5	446	12.0	2	23	202	7	376
6	1358	9.5	1	23	521	8	1184
7	450	10.4	1	23	196	1	417
8	726	7.0	1	18	179	2	525
9	483	6.8	1	22	112	2	355
10	557	7.4	1	19	161	4	485

Note.
[a]Number of items rated by a person in each school.
[b] Number of persons who rated an item.

Table 2.
Percentage of Valid Responses for Each Item across 10 Schools

Item	*Mean %*	*Minimum*	*Maximum*
1	73.9	52.8	89.1
2	15.6	5.4	29.4
3	19.5	9.0	35.7
4	11.6	5.2	19.3
5	42.0	16.5	79.3
6	83.6	72.3	93.1
7	42.2	10.6	70.9
8	16.8	2.7	54.9
9	23.0	14.2	54.9
10	61.5	46.0	75.8
11	24.6	15.6	33.4
12	42.6	3.9	64.7
13	61.5	42.5	75.1
14	44.5	17.8	71.5
15	21.3	10.4	54.5
16	57.2	28.4	88.4
17	8.7	3.1	24.7
18	12.9	5.3	28.7
19	72.3	42.1	92.7
20	13.8	0.7	65.1
21	64.2	13.7	83.1
22	2.0	0.2	5.4
23	1.2	0.2	3.9

Table 3.
Average Item Ratings and Ranks for School 9

Item	Content	n[a]	Mean	Rank
1	*Academic advising services*	255	3.98	8
2	*Personal counseling services*	74	4.07	6
3	*Career planning services*	91	3.74	13
4	*Job placement services*	92	3.68	16
5	*Recreational and intramural programs and services*	84	3.79	11
6	*Library facilities and services*	355	3.90	10
7	*Student health services*	51	4.04	7
8	*Student health insurance program*	13	4.08	5
9	*College-sponsored tutorial services*	95	3.91	9
10	*Financial aid services*	272	3.50	19
11	*Student employment services*	80	3.56	18
12	*Residence hall services and programs*	19	2.89	22
13	*Food services*	244	3.14	21
14	*College-sponsored social activities*	86	3.47	20
15	*Cultural programs*	50	3.70	14
16	*College orientation program*	200	3.69	15
17	*Credit-by-examination program (PEP, CLEP, etc)*	37	4.19	3
18	*Honors programs*	45	4.18	4
19	*Computer services*	284	3.75	12
20	*College mass transit services*	80	3.65	17
21	*Parking facilities and services*	66	2.05	23
22	*Veterans services*	5	4.60	1
23	*Day care services*	2	4.50	2

Note.
[a] Number of Persons Rating the Item

Table 4.
Paired Item Analysis for Item 4 and 5 in School 9

Group[a]	item	n	Pleasability	Mean	Difference[b]
Only	4	59	0.57*	3.61	-0.29
	5	51	1.22	3.90	
Both	4	33	0.79	3.82	0.21
	5	33	0.79	3.61	
Total	4	92	0.65	3.68	-0.11
	5	84	1.05	3.79	
IRT	4	382	1.03	3.85	0.09
	5	382	1.03	3.76	

Note.
[a] Group: only=respondents rating one item only, only=respondents rating both items, total=all respondents rating the item, IRT=all possible respondents predicted by IRT.
[b] Item 4 mean minus Item 5 mean.

Table 5. Difference in Pleasability of Non-overlapping Groups Rating Item-pairs by School

School	# of Pairs	Percent of p <.05
1	246	17
2	249	27
3	250	27
4	253	14
5	249	23
6	251	9
7	200	8
8	253	31
9	242	14
10	244	19

Table 6. Comparisons between IRT and SM Methods in Proximity to Criterion

Item Pair with sample size

# of Persons Rating both Items	*# of Item Pair*	σ_{simple}	σ_{irt}	π_{simple}	π_{irt}	ρ_{simple}	ρ_{irt}
1	2493	0.40	0.31	10%	8%	0.82	0.88
2-9	373	0.65	0.55	18%	13%	0.54	0.64
10-19	241	0.32	0.28	16%	12%	0.85	0.89
20-49	451	0.22	0.18	12%	10%	0.93	0.95
50-99	471	0.16	0.12	10%	8%	0.96	0.98
100-199	380	0.13	0.10	5%	4%	0.98	0.99
200-499	349	0.07	0.06	5%	3%	0.99	1.00
500 or more	112	0.04	0.04	2%	1%	1.00	1.00

Note.
σ_{simple} = Standard deviation of differences between simple mean and criterion mean.
σ_{irt} = Standard deviation of differences between IRT predicted mean and criterion mean.
π_{simple} = Percent of sign disagreement between simple mean and criterion mean.
π_{irt} = Percent of sign disagreement between IRT predicted mean and criterion mean.
ρ_{simple} = Correlation coefficient between simple mean and criterion mean.
ρ_{irt} = Correlation coefficient between IRT predicted mean and criterion mean.

Table 7. Comparison of IRT Predicted Means and Simple Means for School 9 (n=483)

Item	Content	n[a]	Pleasability	IRT	Mean Simple	Char
4	*Job placement services*	92	0.6	3.8	3.7	0..
11	*Student employment services*	80	0.7	3.7	3.6	0.
12	*Residence hall services and programs*	19	0.7	3.2	2.9	0..
3	*Career planning services*	91	0.8	3.9	3.7	0.
17	*Credit-by-examination program (PEP, CLEP, etc)*	37	0.8	4.3	4.2	0.
10	*Financial aid services*	272	0.9	3.5	3.5	0.
18	*Honors programs*	45	0.9	4.2	4.2	0.
19	*Computer services*	284	0.9	3.8	3.8	0.
1	*Academic advising services*	255	1.0	4.0	4.0	0.
6	*Library facilities and services*	355	1.0	3.9	3.9	0.
9	*College-sponsored tutorial services*	95	1.0	3.9	3.9	0.
13	*Food services*	244	1.0	3.1	3.1	0.
16	*College orientation program*	200	1.0	3.6	3.7	0.
5	*Recreational and intramural programs and services*	84	1.1	3.8	3.8	0.
7	*Student health services*	51	1.1	4.0	4.0	0.
14	*College-sponsored social activities*	86	1.1	3.4	3.5	0.
20	*College mass transit services*	80	1.1	3.6	3.7	0.
21	*Parking facilities and services*	66	1.1	2.0	2.0	0.
2	*Personal counseling services*	74	1.2	4.0	4.1	0.
15	*Cultural programs*	50	1.3	3.5	3.7	-0.
8	*Student health insurance program*	13	1.9	3.7	4.1	-0.
22	*Veterans services*	5	2.1	4.4	4.6	-0.
23	*Day care services*	2	4.0	3.5	4.5	-1.

Note.
[a] Number of persons rating the item

BEST COPY AVAILABLE

Table 8. Changes in Rank between IRT Mean and Simple Mean

Change[a]	Frequency	Pecent
No Change	78	33.9
0< change <=2	80	34.8
2< change <=3	37	16.1
3< change <=5	21	9.1
5< change <=10	12	5.2
change > 10	2	0.9
Total	230	100

Note.
[a] Computed as absolute value

Table 9. Changes between IRT Predicted Mean and Simple Mean

Change[a]	Frequency	Percent
No Change	132	57.39
0.1	61	26.52
0.2	22	9.57
0.3	6	2.61
0.4	2	0.87
0.5 or larger	7	3.04
Total	230	100

Note.
[a] Computed as absolute value.

Table 10. Summary Information for Misfit Items

			Overall[a]		Cynics		Neutrals		Pollyannas	
Item	*School*	*fit*	*Mean*	*n*	*Mean*	*n*	*Mean*	*n*	*Mean*	*n*
Overfit Items, fit < .6										
20	3	0.4	2.5	541	1.7	175	2.4	189	3.3	177
23	3	0.3	3.6	39	2.6	13	4.3	11	4.0	15
22	4	0.5	4.0	24	3.8	8	3.8	8	4.4	8
22	5	0.5	3.1	11	2.6	7	4.0	2	4.0	2
23	5	0.2	3.4	7	3.3	4	3.0	1	4.0	2
20	7	0.5	4.0	5	3.0	2	x	0	4.7	3
12	9	0.4	2.9	19	2.1	10	3.4	5	4.3	4
23	9	0	4.5	2	x	0	x	0	4.5	2
Underfit Items, fit > 1.4										
8	1	1.8	3.8	88	3.5	35	3.9	32	4.1	21
18	1	1.9	3.7	20	2.7	7	4.0	5	4.4	8
21	1	1.5	2.8	274	2.1	90	2.7	91	3.5	93
22	2	1.5	4.3	13	3.7	6	4.5	2	5.0	5
18	6	1.5	3.9	147	3.1	49	4.0	53	4.5	45
22	8	2.3	3.0	3	3.0	3	x	0	x	0
21	9	2	2.0	66	1.7	20	1.6	27	3.0	19
9	10	1.5	4.1	85	3.6	21	4.0	33	4.4	31

Note.
[a] All respondents rating the item.

Table 11: Effect of IRT Analysis on Item 4: Job Placement Services

School	Pleasability group	Number of raters	Actual mean	IRT-expected mean	Outfit mean square	Change in rank
	cynics	14	2.71	2.82		
1	neutrals	13	4.0	3.7	1.37	-1
	pollyannas	9	4.1	4.4		
	cynics	12	2.7	2.5		
2	neutrals	13	3.4	3.5	.88	-1
	pollyannas	11	3.9	3.9		
	cynics	17	2.9	2.9		
3	neutrals	16	3.7	3.7	1.09	-1
	pollyannas	17	4.2	4.1		
	cynics	95	3.2	3.3		
4	neutrals	92	3.9	3.9	1.35	-4.5
	pollyannas	73	4.4	4.3		
	cynics	27	3.3	3.1		
5	neutrals	20	3.7	3.9	1.21	-2
	pollyannas	22	4.5	4.4		
	cynics	72	2.5	2.5		
6	neutrals	56	3.4	3.4	1.08	+1
	pollyannas	52	4.1	4.0		
	cynics	16	3.8	3.5		
7	neutrals	15	3.8	4.0	.99	-2
	pollyannas	13	4.3	4.4		
	cynics	42	3.1	3.1		
8	neutrals	21	3.5	3.9	1.04	-3
	pollyannas	14	4.4	4.4		
	cynics	44	3.2	3.2		
9	neutrals	31	4.0	3.9	.90	-6
	pollyannas	17	4.5	4.4		
	cynics	14	3.1	3.1		
10	neutrals	16	3.9	3.8	.82	-2
	pollyannas	8	4.2	4.4		

Figure 1. Distribution of the Measured Pleasability for School 9 (*n=483*)

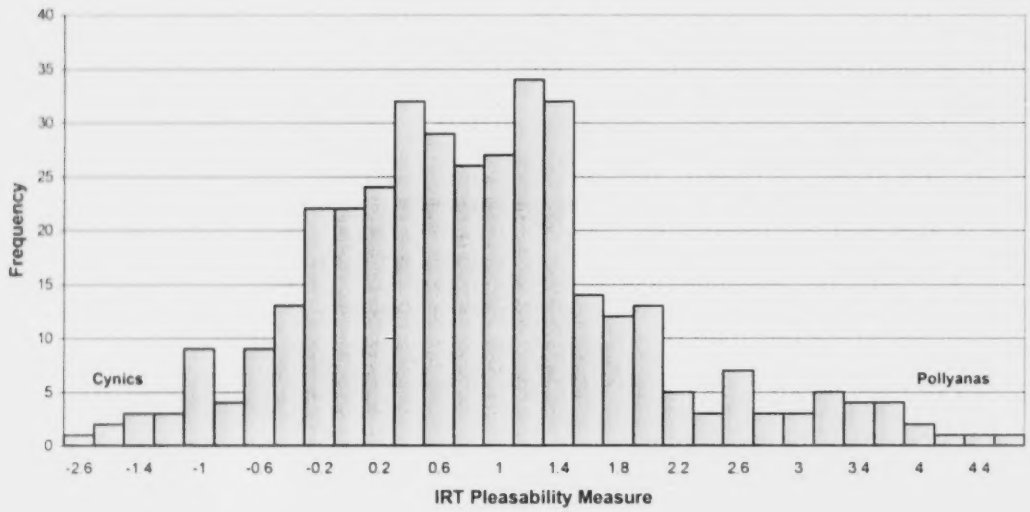

Figure 2. Scatter Plot of Reliability of Pleasability Measure

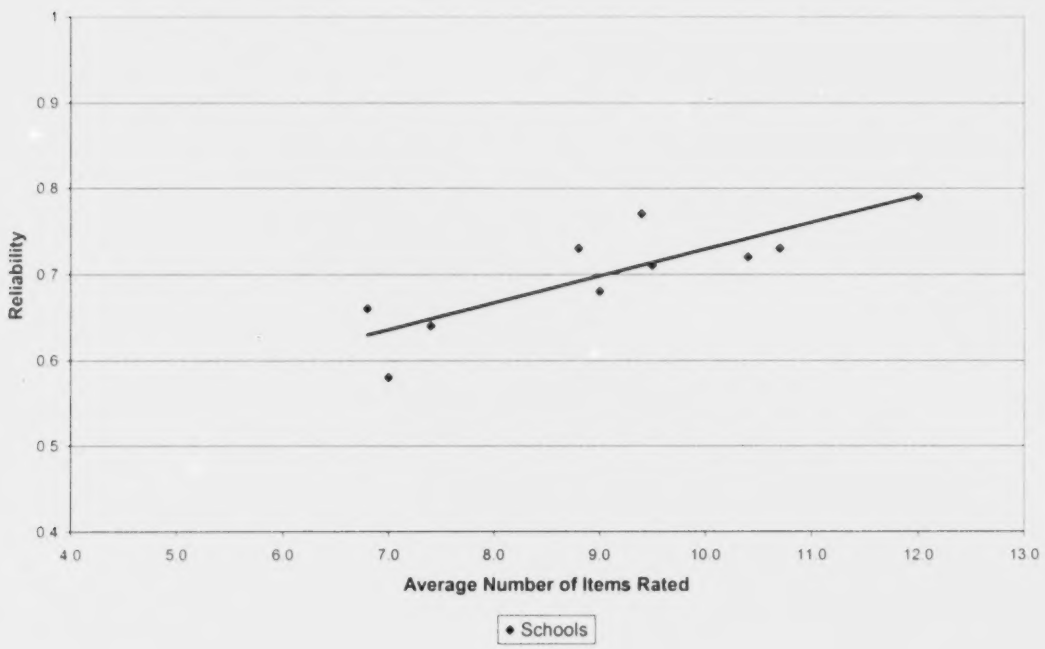

Figure 3. Fit Statistics for the 23 items across 10 Schools

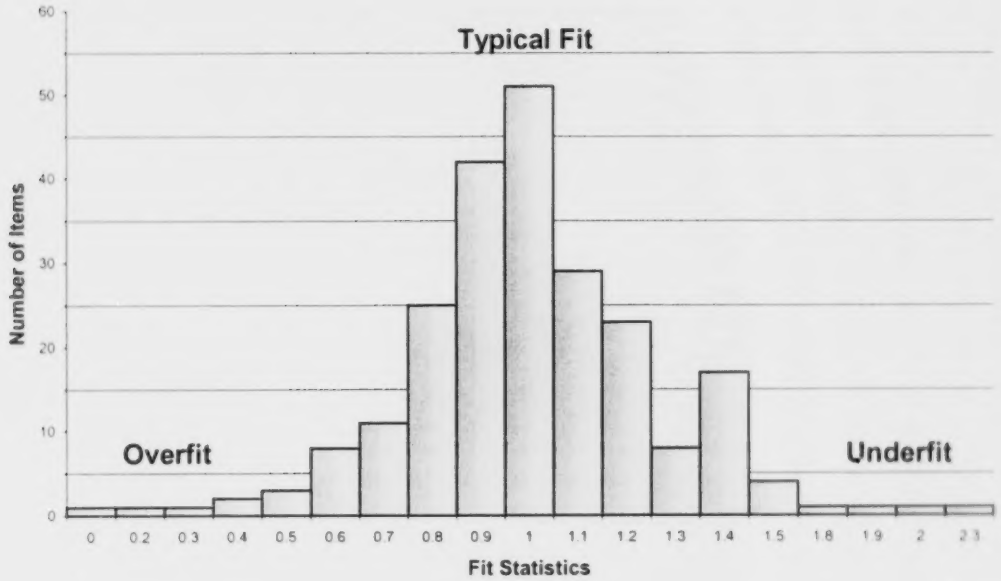

Figure 4. Plots of Expected and Observed Means for Overfitted and Underfitted Items